75th MORSS CD Cover Slide

712CD

For office use only 41205

UNCLASSIFIED DISCLOSURE FORM CD Presentation

12-14 June 2007, at US Naval Academy, Annapolis, MD

Please complete this form 712CD as your cover page to your electronic briefing submission to the MORSS CD. Do not fax to the MORS office.

Author Request **(To be completed by applicant) - The following author(s) request authority to disclose the following presentation in the MORSS Final Report, for inclusion on the MORSS CD and/or posting on the MORS web site.**

Name of Principal Author and all other author(s): James C. Spall

Principal Author's Organization and address:
The Johns Hopkins University
Applied Physics Laboratory
Laurel, MD 20723-6099

Phone: 240-228-4960

Fax: 240-228-8110

Email: james.spall@jhuapl.edu.

Original title on 712 A/B: Stochastic Optimization and the Simultaneous Perturbation Method

Revised title: No change to above title

Presented in: ***Tutorial Session***

This presentation is believed to be:
Distribution A: UNCLASSIFIED AND APPROVED FOR PUBLIC RELEASE

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JUN 2007	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Stochastic Optimization and the Simultaneous Perturbation Method	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The Johns Hopkins University Applied Physics Laboratory Laurel, MD 20723-6099**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM202526. Military Operations Research Society Symposium (75th) Held in Annapolis, Maryland on June 12-14, 2007, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**37**	

STOCHASTIC OPTIMIZATION AND THE SIMULTANEOUS PERTURBATION ALGORITHM

MORS Symposium, 13 June 2007

James C. Spall
The Johns Hopkins University
Applied Physics Laboratory (JHU/APL)

james.spall@jhuapl.edu

Organization

- A. Problem setting
- B. SPSA algorithm
- C. Theoretical foundation
- D. Practical guidelines–MATLAB code
- E. Numerical example
- F. Adaptive simultaneous perturbation method
- G. Extensions and further results

Additional information available at ***www.jhuapl.edu/SPSA*** (reference list, background articles, MATLAB code, and video)

A. PROBLEM SETTING

- Focus here is on ***stochastic*** search and optimization:

> **A. Random noise in input information (e.g., noisy measurements of loss function)**
>
> **— and/or —**
>
> **B. Injected randomness (Monte Carlo) in choice of algorithm iteration magnitude/direction**

- Contrasts with deterministic methods
  - — E.g., steepest descent, Newton-Raphson, etc.
  - — Assume perfect information about $L(\theta)$ (and its gradient)
  - — Search magnitude/direction deterministic at each iteration
- Injected randomness (B) in search magnitude/direction can offer benefits in efficiency and robustness
  - — E.g., Capabilities for global (vs. local) optimization

Some Popular Stochastic Search and Optimization Techniques

- Random search
- Stochastic approximation
  - Robbins-Monro and Kiefer-Wolfowitz
  - SPSA
  - NN backpropagation
  - Infinitesimal perturbation analysis
  - Recursive least squares
  - Many others
- Simulated annealing
- Evolutionary computation and genetic algorithms
- Reinforcement learning
- Markov chain Monte Carlo (MCMC)
- Etc.

Baseline Problem Setting for SPSA Algorithm

- Consider standard minimization setting, i.e., find root θ^* to

$$\boldsymbol{g}(\theta) = \frac{\partial L(\theta)}{\partial \theta} = 0$$

  where $L(\theta)$ is scalar-valued loss function to be minimized and θ is p-dimensional vector
- Assume only (possibly noisy) measurements of $L(\theta)$ available
  - No direct measurements of $\boldsymbol{g}(\theta)$ used, as are required in stochastic gradient methods
- Noisy measurements of $L(\theta)$ in areas such as Monte Carlo simulation, real-time control/estimation, etc.
- Interested in $p > 1$ setting (including $p >> 1$)

B. SPSA ALGORITHM

- Let $\hat{\boldsymbol{g}}_k(\theta)$ denote SP estimate of $\boldsymbol{g}(\theta)$ at kth iteration
- Let $\hat{\theta}_k$ denote estimate for θ^* at kth iteration
- SPSA algorithm has form

$\hat{g}_k(\hat{\theta}_k)$ is critical !

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$$

where $\{a_k\}$ is nonnegative gain sequence

- Generic iterative form above is standard in SA; stochastic analogue to steepest descent
- Under conditions, $\hat{\theta}_k \to \theta^*$ in some stochastic sense as $k \to \infty$

Computation of $\hat{g}_k(\bullet)$ (Heart of SPSA)

- Let Δ_k be vector of p independent random variables at kth iteration

$$\Delta_k = \left[\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}\right]^T$$

- Δ_k typically generated by Monte Carlo
- Let $\{c_k\}$ be sequence of positive scalars
- For iteration $k \rightarrow k+1$, take measurements at design levels: $\hat{\theta}_k \pm c_k \Delta_k$

$$y(\hat{\theta}_k + c_k\Delta_k) = L(\hat{\theta}_k + c_k\Delta_k) + \varepsilon_k^{(+)}$$
$$y(\hat{\theta}_k - c_k\Delta_k) = L(\hat{\theta}_k - c_k\Delta_k) + \varepsilon_k^{(-)}$$

  where $\varepsilon_k^{(\pm)}$ are measurement noise terms

- Common special case is when $\varepsilon_k^{(\pm)} = 0 \; \forall \, k$ (e.g., system identification with perfect measurements of the likelihood function)

Computation of $\hat{\boldsymbol{g}}_k(\bullet)$ (cont'd)

- The standard SP form for $\hat{\boldsymbol{g}}_k(\bullet)$:

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \dfrac{y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{k1}} \\ \vdots \\ \dfrac{y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{kp}} \end{bmatrix}$$

- Note that $\hat{\boldsymbol{g}}_k(\bullet)$ only requires ***two*** measurements of $L(\bullet)$ ***independent*** of ***p***

- Above SP form contrasts with standard finite-difference approximations taking $2p$ (or p+1) measurements

- Intuitive reason why $\hat{\boldsymbol{g}}_k(\bullet)$ is appropriate is that $E[\hat{\boldsymbol{g}}_k(\hat{\theta}_k) \,|\, \hat{\theta}_k] \approx \boldsymbol{g}(\hat{\theta}_k)$; formalized in Section C

Essential Conditions for SPSA

- To use SPSA, there are regularity conditions on $L(\theta)$, choice of Δ_k, the gain sequences $\{a_k\}$, $\{c_k\}$, and the measurement noise
  - Sections 7.3 and 7.4 of *ISSO* present essential conditions
- Roughly speaking the conditions are:

  A. ***L*(θ) *smoothness:*** $L(\theta)$ is thrice differentiable function (can be relaxed—see Section 7.3 of *ISSO*)

  B. ***Choice of*** Δ_k ***distribution:*** For all k, Δ_k has independent components, symmetrically distributed around 0, and $E(\Delta_{ki}^2) < \infty$, $E(\Delta_{ki}^{-2}) < \infty$
  - Bounded inverse moments condition is critical (***excludes*** Δ_{ki} being normally or uniformly distributed)
  - Symmetric Bernoulli $\Delta_{ki} = \pm 1$ (prob = ½ for each outcome) is allowed; asymptotically optimal (see Section G or Section 7.7 of *ISSO*)

Essential Conditions for SPSA (cont'd)

C. ***Gain sequences:*** standard SA conditions:

$$a_k, c_k > 0, a_k, c_k \to 0 \text{ as } k \to \infty$$

$$\sum_{k=0}^{\infty} a_k = \infty, \sum_{k=0}^{\infty} \left(\frac{a_k}{c_k} \right)^2 < \infty$$

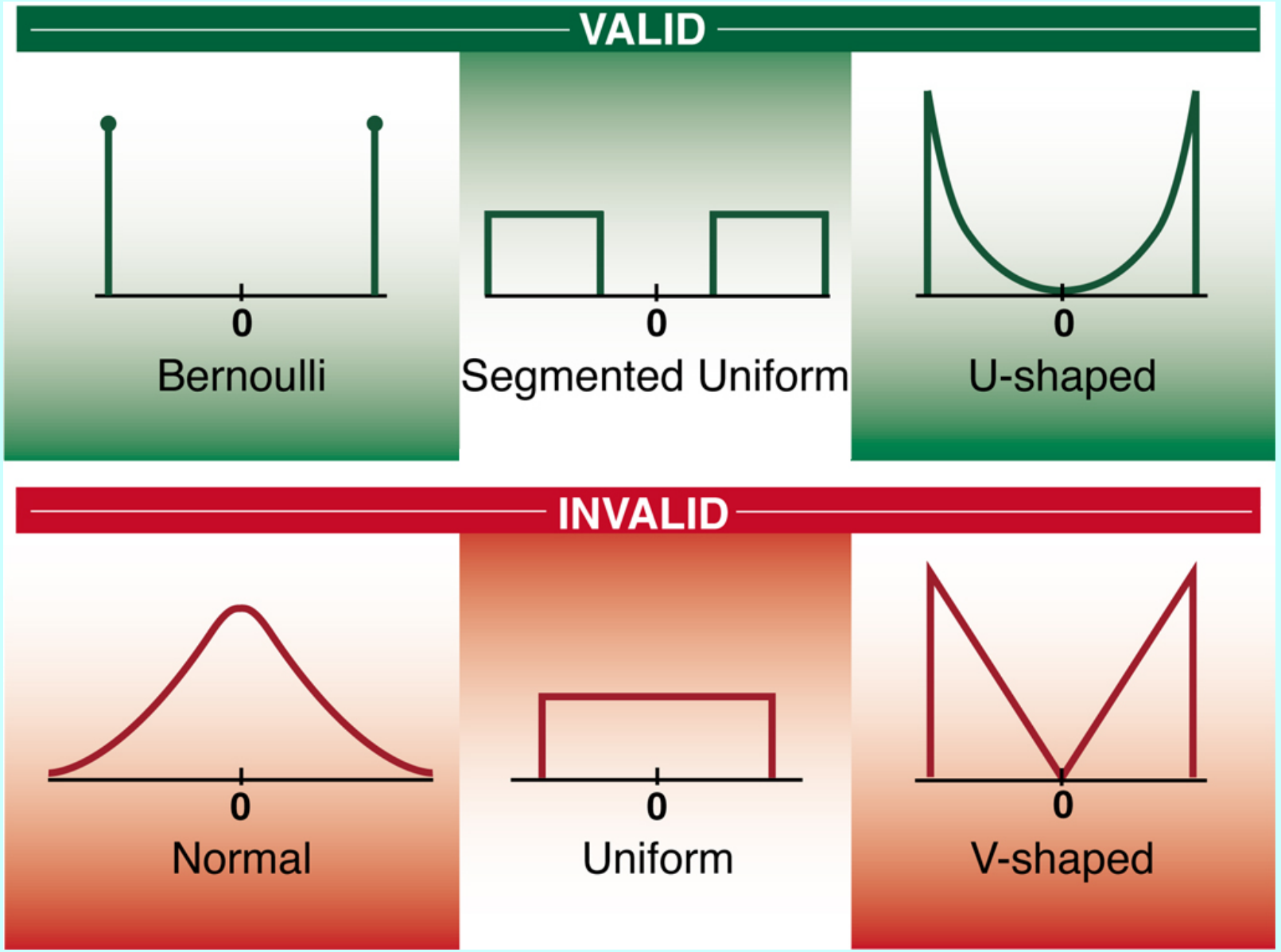
(better to violate some of these gain conditions in certain practical problems; e.g., nonstationary tracking and control where $a_k = a > 0$, $c_k = c > 0$ $\forall$ k, i)

D. ***Measurement Noise:*** Martingale difference

$$E[\varepsilon_k^{(+)} - \varepsilon_k^{(-)} \,|\, \hat{\theta}_k, \Delta_k] = 0$$

$\forall$ k sufficiently large. (Noises ***not*** required to be independent of each other or of current/previous $\hat{\theta}_k$ and Δ_k values.) ***Alternative*** condition (no martingale mean 0 assumption needed) is that $\varepsilon_k^{(\pm)}$ be bounded $\forall$ k

Valid and Invalid Perturbation Distributions



C. THEORETICAL FOUNDATION

Three Questions

Question 1: Is $\hat{\boldsymbol{g}}_k(\bullet)$ a valid estimator for $\boldsymbol{g}(\bullet)$?

Answer: Yes, under modest conditions.

Question 2: Will the algorithm converge to θ^*?

Answer: Yes, under reasonable conditions.

Question 3: Do savings in data/iteration lead to a corresponding savings in converging to optimum?

Answer: Yes, under reasonable conditions.

Near Unbiasedness of $\hat{\boldsymbol{g}}_k(\bullet)$

- SPSA stochastic analogue to deterministic algorithms if $\hat{\boldsymbol{g}}_k(\theta)$ is "on average" same as $\boldsymbol{g}(\theta)$ for any θ
- Suppressing iteration index k, mth component of $\hat{\boldsymbol{g}}(\theta)$ is:

$$\hat{g}_m(\theta) = \frac{L(\theta + c\Delta) - L(\theta - c\Delta)}{2c\Delta_m} + noise$$

$$\approx \frac{L(\theta) + c\boldsymbol{g}(\theta)^T\Delta - L(\theta) - (-c\boldsymbol{g}(\theta)^T\Delta)}{2c\Delta_m} + noise$$

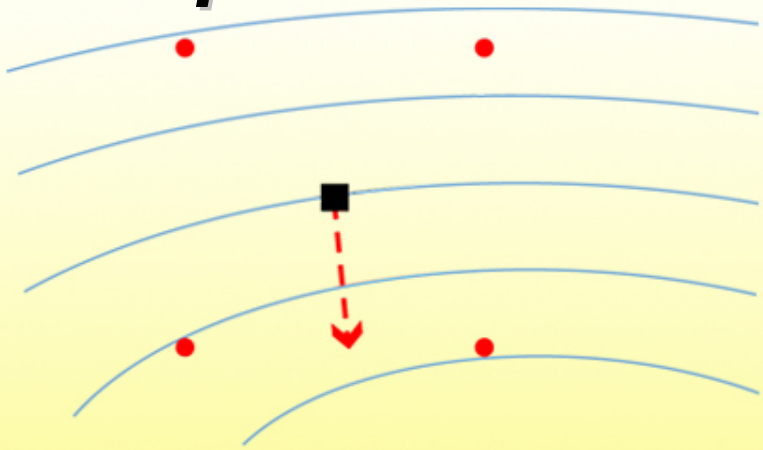
$$= \frac{\sum_i g_i(\theta)\Delta_i}{\Delta_m} + noise$$

$$= g_m(\theta) + \sum_{i \neq m} g_i(\theta)\frac{\Delta_i}{\Delta_m} + noise$$

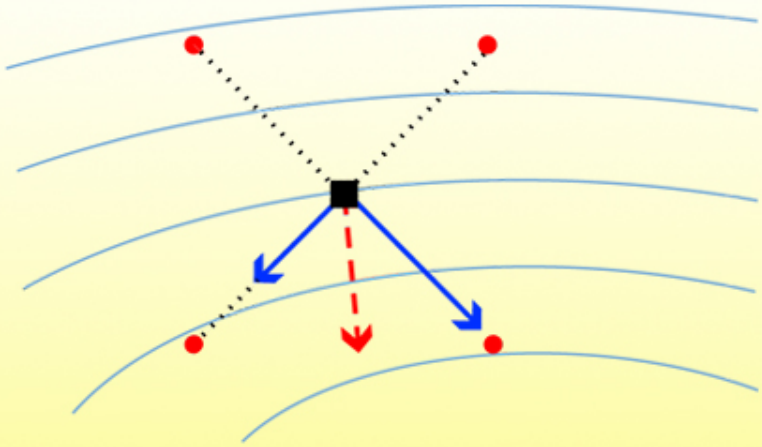
- With $E(\Delta_i / \Delta_m) = 0$ we have for any m:

$$\boxed{E[\hat{g}_m(\theta)] = g_m(\theta) + negligible\ terms}$$

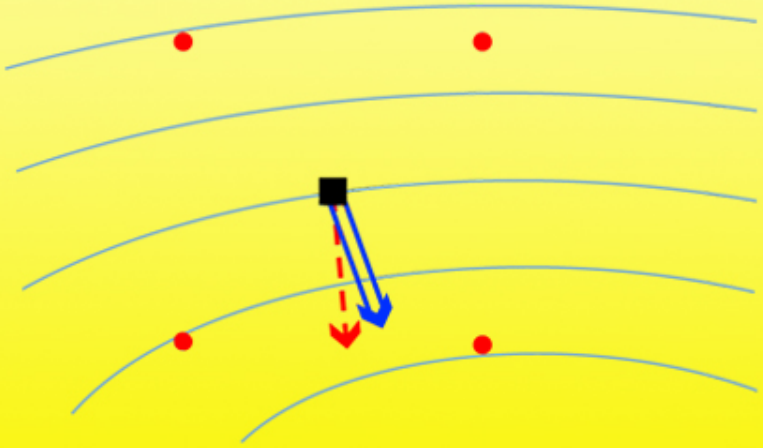
Illustration of Near-Unbiasedness for $\hat{g}_k(\bullet)$ with $p = 2$ and Bernoulli Perturbations



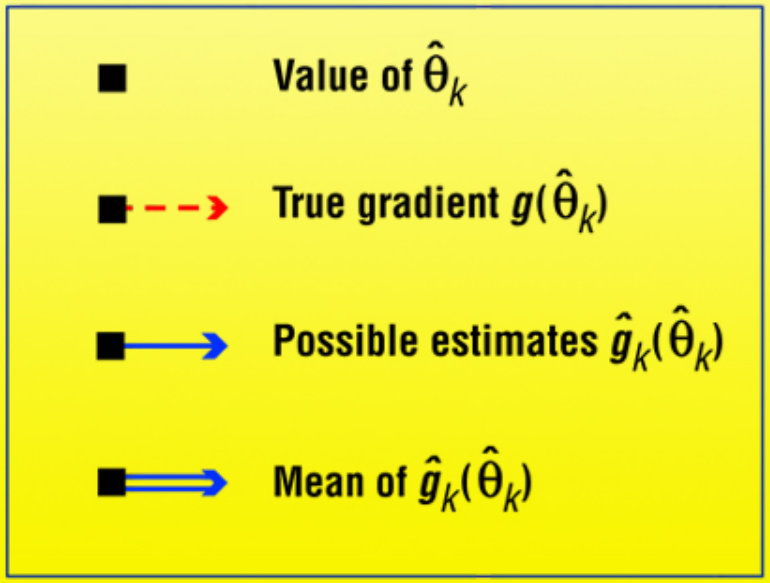
(a) True gradient and four possible sample points around $\hat{\theta}_k$



(b) Two possible search directions and magnitudes $\hat{g}_k(\hat{\theta}_k)$



(c) Mean of the two possible $\hat{g}_k(\hat{\theta}_k)$ values: $E[\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k] \approx g(\hat{\theta}_k)$



Theoretical Basis (Sects. 7.3 – 7.4 of *ISSO*)

- Under appropriate regularity conditions (e.g., $E(\Delta_{ki}^{-2}) < \infty, L(\theta)$ thrice continuously differentiable, $\varepsilon_k^{(\pm)}$ is martingale difference noise, etc.), we have:
- ***Near Unbiasedness***

$$E[\hat{\boldsymbol{g}}_k(\hat{\theta}_k) \,|\, \hat{\theta}_k] = \boldsymbol{g}(\hat{\theta}_k) + O(c_k^2) \text{ a.s.}$$

  where $c_k \to 0$

- ***Convergence:***

$$\hat{\theta}_k \to \theta^* \text{ a.s. as } k \to \infty$$

- ***Asymptotic Normality:***

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist.}} N(\mu, \Sigma), \quad 0 < \beta \le 2/3$$

  where μ, Σ, and β depend on SA gains, Δ_k distribution, and shape of $L(\theta)$

Efficiency Analysis

- Can use asymptotic normality to analyze relative efficiency of SPSA and FDSA (Spall, 1992; Sect. 7.4 of *ISSO*)
- Analogous to SPSA asymptotic normality result, FDSA is also asymptotically normal (Chap. 6 of *ISSO*)

> The critical cost in comparing relative efficiency of SPSA and FDSA is number of loss function measurements $y(\bullet)$, not number of iterations per se

- Loss function measurements represent main cost (by far)—other costs are trivial
- Full efficiency story is fairly complex—see Section 7.4 of *ISSO* and references therein

Efficiency Analysis (cont'd)

- Will compare SPSA and FDSA by looking at relative mean square error (MSE) of θ estimate
- Consider relative MSE for same no. of measurements, *n* (***not*** same no. of iterations). Under regularity conditions above:

$$\frac{E\left(\left\|\hat{\theta}_{SPSA,n}-\theta^*\right\|^2\right)}{E\left(\left\|\hat{\theta}_{FDSA,n}-\theta^*\right\|^2\right)} \to \frac{1}{p^{\beta}},\ 0<\beta\le 2/3$$

☺

as $n \to \infty$

- Equivalently, to achieve ***same asymptotic MSE***

$$\frac{no.\ meas.\ y(\theta)\ in\ SPSA}{no.\ meas.\ y(\theta)\ in\ FDSA}=\frac{1}{p}$$

☺ ☺

- Results ☺ and ☺ ☺ are main theoretical results

Paraphrase of ☺ ☺ above:

- SPSA and FDSA converge in same number of iterations despite *p*-fold savings in cost/iteration for SPSA

— or —

- One properly generated simultaneous random change of all variables in a problem contains as much information ***for optimization*** as a full set of one-at-a-time changes of each variable

D. PRACTICAL GUIDELINES AND MATLAB CODE

- The code below implements SPSA iterations $k = 1,2,\ldots,n$
  - Initialization for program variables `theta`, `alpha`, etc. not shown since that can be handled in numerous ways (e.g., file read, direct inclusion, input during execution)
  - Δ_k elements are generated by Bernoulli ±1
  - Program calls external function `loss` to obtain $y(\theta)$ values
- Simple enhancements possible to increase algorithm stability and/or speed convergence
  - Check for simple constraint violation (shown at bottom of sample code)
  - Reject iteration $k \to k+1$ if $y(\hat{\theta}_{k+1})$ is too much greater than $y(\hat{\theta}_k)$ (requires extra loss measurement per iteration)
  - Reject iteration $k \to k+1$ if $\left\|\hat{\theta}_{k+1} - \hat{\theta}_k\right\|$ is too large (does not require extra loss measurement)

Matlab Code

```
for k=1:n
     ak=a/(k+A)^alpha;
     ck=c/k^gamma;
     delta=2*round(rand(p,1))-1;
     thetaplus=theta+ck*delta;
     thetaminus=theta-ck*delta;
     yplus=loss(thetaplus);
     yminus=loss(thetaminus);
     ghat=(yplus-yminus)./(2*ck*delta);
     theta=theta-ak*ghat;
end
theta
```

If maximum and minimum values on elements of `theta` can be specified, say `thetamax` and `thetamin`, then two lines can be added below `theta` update line to impose constraints:

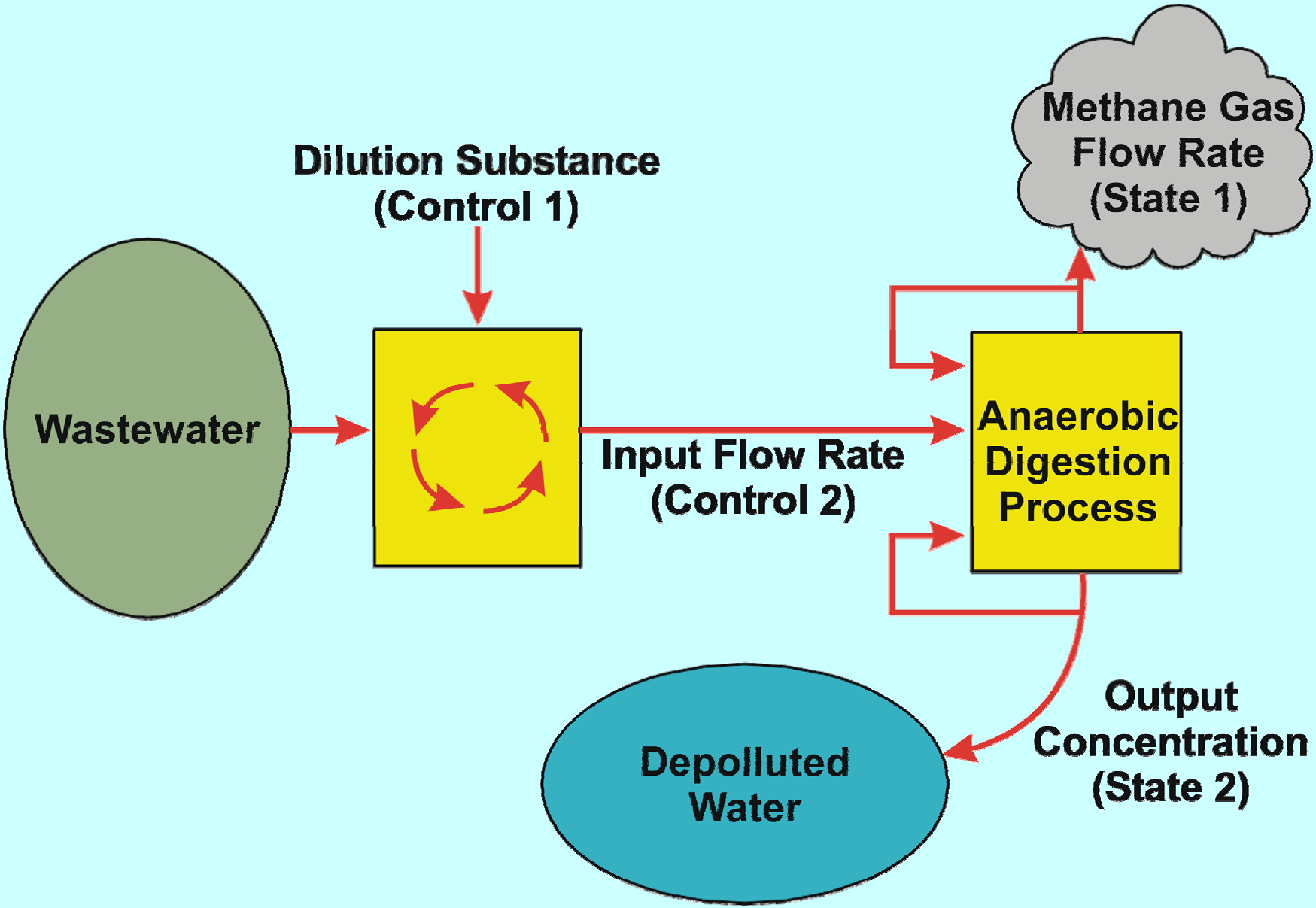
```
     theta=min(theta,thetamax);
     theta=max(theta,thetamin);
```

E. APPLICATION OF SPSA

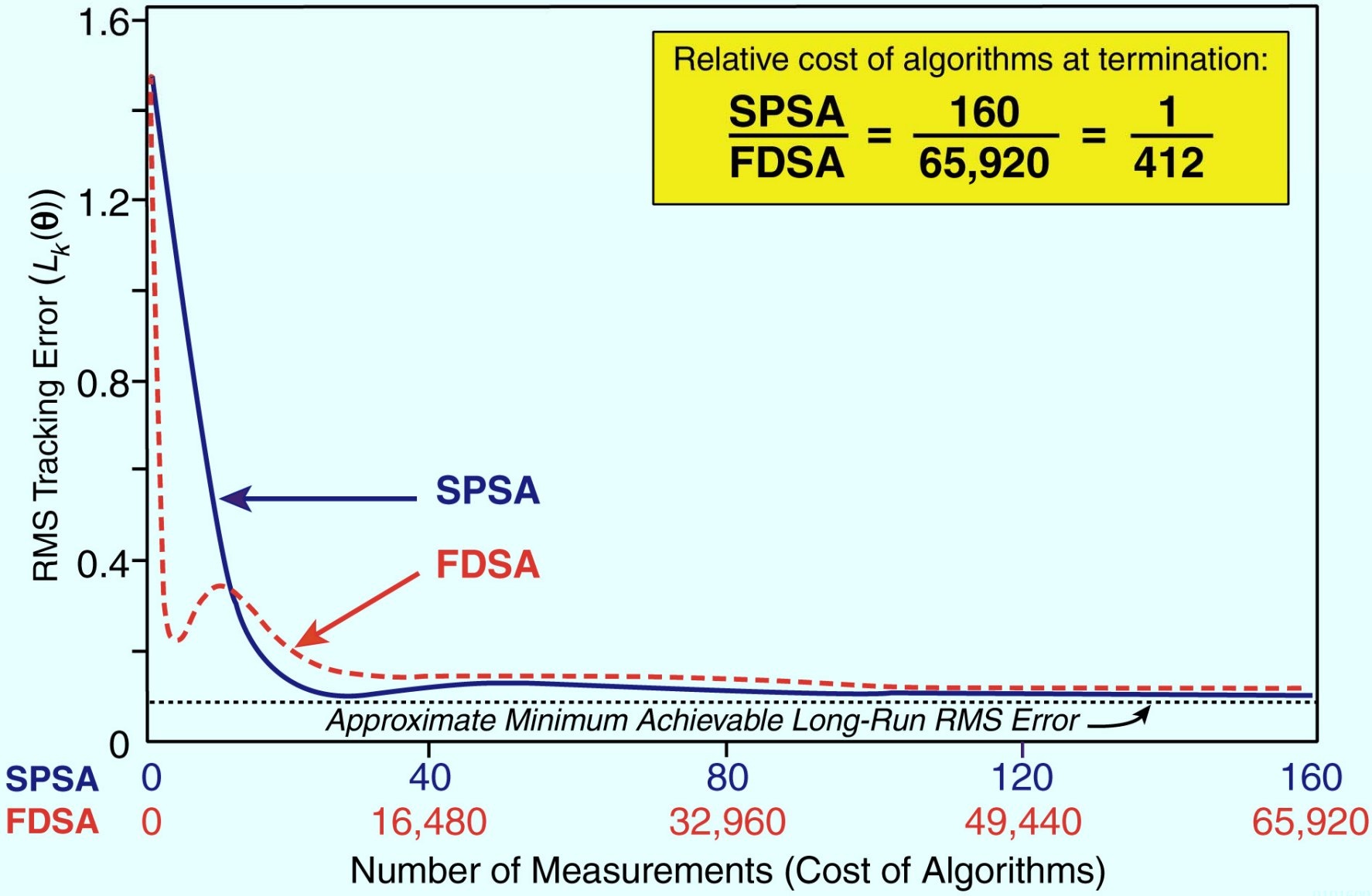
- Numerical Study: SPSA vs. FDSA
- Consider problem of developing neural net controller (wastewater treatment plant where objectives are clean water ***and*** methane gas production)
- Neural net is function approximator that takes current information about the state of system and produces control action
- $L_k(\theta)$ = tracking error,

  θ = neural net weights
- Need to estimate θ in real-time; used nondecaying $a_k = a$, $c_k = c$ due to nonstationary dynamics
- $p = \dim(\theta) = 412$
- More information in Example 7.4 of *ISSO*

Wastewater Treatment System



RMS Error for Controller in Wastewater Treatment Model



F. ADAPTIVE SIMULTANEOUS PERTURBATION METHOD

- Standard SPSA exhibits common "1st-order" behavior
  - Sharp initial decline
  - Slow convergence in final phase
  - Sensitivity to units/scaling for elements of θ
- "2nd-order" form of SPSA exists for speeding convergence, especially in final phase (analogous to Newton-Raphson)
  - Adaptive simultaneous perturbation (ASP) method (details in Section 7.8 of *ISSO*)
- ASP based on adaptively estimating Hessian matrix

$$\boldsymbol{H}(\theta) \equiv \frac{\partial^2 L(\theta)}{\partial \theta \partial \theta^T}$$

- Addresses long-standing problem of finding "easy" method for Hessian estimation
- Also has uses in nonoptimization applications (e.g., Fisher information matrix in Subsection 13.3.5 of *ISSO*)

Overview of ASP

- ASP applies in either

  (i) Standard SPSA setting where only $L(\theta)$ measurements are available (as considered earlier) ("2SPSA" algorithm)

  — or —

  (ii) Stochastic gradient (SG) setting where $L(\theta)$ and $\boldsymbol{g}(\theta)$ measurements are available ("2SG" algorithm)

- Advantages of 2nd-order approach
  - — Potential for speedier convergence
  - — Transform invariance (algorithm performance unaffected by relative magnitude of θ elements)
- Transform invariance is unique to 2nd-order algorithms
  - — Allows for arbitrary scaling of θ elements
  - — Implies ASP automatically adjusts to chosen units for θ

Cost of Implementation

- For any p, the cost per iteration of ASP is

Four loss measurements for 2SPSA

— or —

Three gradient measurements for 2SG

- Above costs for ASP compare very favorably with previous methods:

  $O(p^2)$ loss measurements ($y(\bullet)$) per iteration in FDSA setting (e.g., Fabian, 1971)

  $O(p)$ gradient measurements per iteration in SG setting (e.g., Ruppert, 1985)

- If gradient/Hessian averaging or $y(\bullet)$-based iterate blocking is used, then additional measurements needed per iteration

Efficiency Analysis for ASP

- Can use asymptotic normality of 2SPSA and 2SG to compare asymptotic RMS errors (as in basic SPSA) against ***best possible*** asymptotic RMS of SPSA and SG, say RMS^*_{SPSA} and RMS^*_{SG}
- ***2SPSA***: With $a_k = 1/k$ and $c_k = c\,/k^{1/6}$ ($k \geq 1$)

$$\frac{\text{RMS of 2SPSA}}{RMS^*_{SPSA}} < 2 \quad \forall\, c > 0$$

- ***2SG***: With $a_k = 1/k$ and any valid c_k

$$\frac{\text{RMS of 2SG}}{RMS^*_{SG}} = 1$$

- ***Interpretation***: 2SPSA (with $a_k = 1/k$) does almost as well as ***unobtainable*** best SPSA; RMS error differs by < factor of 2
- 2SG (with $a_k = 1/k$) does as well as the analytically optimal SG (rarely available)

G. EXTENSIONS AND FURTHER RESULTS

- There are variations and enhancements to "standard" SPSA of Section B
- Section 7.7 of *ISSO* discusses:
  - (i) Enhanced convergence through gradient averaging/smoothing
  - (ii) Constrained optimization
  - (iii) Optimal choice of Δ_k distribution
  - (iv) One-measurement form of SPSA
  - (v) Global optimization
  - (vi) Noncontinuous (discrete) optimization

(i) Gradient Averaging and Gradient Smoothing

- These approaches may yield improved convergence in some cases
- In ***gradient averaging*** $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ is simply replaced by the average of several (say, q) SP gradient estimates
  - This approach uses $2q$ values of $y(\bullet)$ per iteration
  - Spall (1992) establishes theoretical conditions for when this is advantageous, i.e., when lower MSE compensates for greater per-iteration cost ($2q$ vs. 2, q >1)
  - Essentially, beneficial in a high-noise environment (consistent with intuition!)
- In ***gradient smoothing***, gradient estimates averaged across iterations according to scheme that carefully balances past estimates with current estimate
  - Analogous to "momentum" in neural net/backpropagation literature

(ii) Constrained Optimization

- Most practical problems involve constraints on θ
- Numerous possible ways to treat constraints (simple constraints discussed in Section D)
- One approach based on ***projections*** (exploits well-known Kuhn-Tucker framework)
- Projection approach keeps $\hat{\theta}_k$ and $\hat{\theta}_k \pm c_k \Delta_k$ in valid region for all *k* by projecting $\hat{\theta}_k$ into a region ***interior*** to the valid region
  - Desirable in real systems to keep $\hat{\theta}_k \pm c_k \Delta_k$ (in addition to $\hat{\theta}_k$) inside valid region to ensure physically achievable solution while iterating
- ***Penalty functions*** are general approach that may be easier to use than projections
  - However, penalty functions require care for efficient implementation

(iii) Optimal Choice of Δ_k Distribution

- Sections 7.3 and 7.4 of *ISSO* discuss sufficient conditions for Δ_k distribution (see also Sections B and C here)
  - These conditions guide user since user typically has full control over distribution
  - Uniform and normal distributions do ***not*** satisfy conditions
- Asymptotic distribution theory shows that ***symmetric Bernoulli*** distribution is asymptotically optimal
  - Optimal in both an MSE and nearness-probability sense
  - Symmetric Bernoulli is trivial to generate by Monte Carlo
- Symmetric Bernoulli seems optimal in many practical (***finite-sample***) problems
  - One exception mentioned in Section 7.7 of *ISSO* (robot control problem): segmented uniform distribution

(iv) One-Measurement SPSA

- Standard SPSA use two loss function measurements/iteration
- ***One-measurement*** SPSA based on gradient approximation:

$$\hat{\boldsymbol{g}}_k(\hat{\theta}_k) = \begin{bmatrix} \dfrac{y(\hat{\theta}_k + c_k\Delta_k)}{c_k\Delta_{k1}} \\ \vdots \\ \dfrac{y(\hat{\theta}_k + c_k\Delta_k)}{c_k\Delta_{kp}} \end{bmatrix}$$

- As with two-measurement SPSA this form is unbiased estimate of $\boldsymbol{g}(\hat{\theta}_k)$ to within $O(c_k^2)$
- Theory shows standard two-measurement form generally preferable in terms of total measurements needed for effective convergence
  - ***However***, in ***some*** settings, one-measurement form is preferable
  - One such setting: control problems with significant nonstationarities

(v) Global Optimization

- SPSA has demonstrated significant effectiveness in ***global optimization*** where there may be multiple (local) minima
- One approach is to inject Gaussian noise to right-hand side of standard SPSA recursion:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\theta}_k) + b_k \boldsymbol{w}_k \qquad (*)$$

  where $b_k \to 0$ and $\boldsymbol{w}_k \sim N(\boldsymbol{0}, \boldsymbol{I}_{p\times p})$

- Injected noise $\boldsymbol{w}_k$ generated by Monte Carlo
- Eqn. (*) has theoretical basis for formal convergence (Section 8.4 of *ISSO*)

(v) Global Optimization (Cont'd)

- Recent results show that $b_k = 0$ is sufficient for global convergence in many cases (Section 8.4 of *ISSO*)
  - ***No injected noise*** needed for global convergence
  - Implies ***standard SPSA*** is global optimizer under appropriate conditions
- Numerical demo on some tough global problems with many local minima yield global solution
  - Neither genetic algorithms nor simulated annealing able to find global minima in test suite
  - No guarantee of analogous relative behavior on other problems
- Regularity conditions for global convergence of SPSA difficult to check

(vi) Noncontinuous (Discrete) Optimization

- Basic SPSA framework for $L(\theta)$ differentiable in θ
- Many important problems have elements in θ taking only discrete (e.g., integer) values
- There have been extensions to SPSA to allow for discrete θ
  - Brief discussion in Section 7.7 of *ISSO*; see also references at SPSA Web site
- SP estimate $\hat{\boldsymbol{g}}_k(\hat{\theta}_k)$ produces descent information although gradient not defined
- Key issue in implementation is to control iterations $\hat{\theta}_k$ and perturbations $\hat{\theta}_k \pm c_k\Delta_k$ to ensure they are valid θ values

Contact and Other Information

- Contact: James C. Spall

  james.spall@jhuapl.edu

  240-228-4960

- SPSA web site

  www.jhuapl.edu/SPSA

- Additional relevant information at site for related book *Introduction to Stochastic Search and Optimization*

  www.jhuapl.edu/ISSO

- Tutorial paper (available at SPSA web site):

  Spall, J. C. (1998), "An Overview of the Simultaneous Perturbation Method for Efficient Optimization," *Johns Hopkins APL Technical Digest*, vol. 19, pp. 482–492.